Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique

Guillaume Surroca, Philippe Lemoisson, Clement Jonquet, Stefano A. Cerri

Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique

Guillaume Surroca[1], Philippe Lemoisson[1,2],
Clément Jonquet[1], Stefano A. Cerri[1]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)
Université Montpellier 2 & CNRS
prenom.nom@lirmm.fr
[2] UMR Territoires, Environnement, Télédétection et Information Spatiale, CIRAD
Montpellier, France
philippe.lemoisson@cirad.fr

Résumé : Avec le Web 2.0, les utilisateurs, devenus contributeurs, ont pris une place centrale dans les processus de consommation et de production de connaissances ; cependant la paternité des contributions est souvent perdue lors de l'indexation de l'information. VIEWPOINTS est un formalisme de représentation des connaissances centré sur le point de vue individuel, humain ou artificiel. Nous considérons trois types d'objets de connaissance : les *documents* (supports), les *agents* (émetteurs) et les *topics* (descripteurs). Un *viewpoint* émis par un *agent* exprime son opinion sur la proximité entre deux objets. Les *viewpoints* permettent de définir et de calculer une distance entre objets qui évolue au fil des interactions (requêtes et retours d'utilisation) et de l'ajout de nouveaux *viewpoints*. Un prototype de moteur de recherche pour des données de publications scientifiques tirées de HAL-LIRMM montre comment VIEWPOINTS peut faire émerger, de façon transparente, une intelligence collective à partir des interactions des utilisateurs contributeurs.

Mots-clés : représentation des connaissances, construction collaborative de connaissances, indexation et recherche d'information, découverte interactive de connaissances, sérendipité, ingénierie des connaissances centrée utilisateurs, distance sémantique, Web 2.0.

1 Introduction

Le Web d'aujourd'hui est un espace de profusion de l'information où les échanges informels des réseaux sociaux cohabitent avec des jeux de données structurées pouvant alimenter des traitements automatiques (confrontation, intégration, raisonnement). C'est d'abord un espace où interagissent les humains, notamment depuis l'avènement du Web 2.0, à la fois en tant que producteurs/consommateurs de ressources, mais aussi en tant que commentateurs sur ces ressources. C'est aussi un espace peuplé d'agents artificiels invisibles, en charge notamment de tâches de fouille, de présentation, de traduction, d'indexation, etc. C'est donc un espace de partage de connaissances, où collaborent des agents humains et des agents artificiels, avec des connaissances collectives stabilisées dans les schémas du Web sémantique (vocabulaires, ontologies) [4] et des contributions individuelles spontanées du Web 2,0 (réseaux sociaux, recommandations) [16]. Face à cette complexité, les internautes sont confrontés à trois questions récurrentes :

- Comment trouver des documents ou des données dignes de confiance sur un sujet particulier ?

- Comment trouver les bonnes personnes pour échanger, argumenter et capitaliser sur un sujet particulier ?
- Peut-on faciliter les processus d'agrégation qui feront émerger de nouvelles connaissances au sein de la communauté ?

Partant de ce constat, notre travail a consisté à : (i) chercher un formalisme de représentation des connaissances qui puisse capturer les contributions individuelles tout en tissant des liens avec les schémas du Web sémantique ; (ii) chercher à proposer des protocoles régissant de façon transparente la recherche d'information. Ce document présente l'approche VIEWPOINTS et son formalisme, ainsi qu'un prototype permettant d'expérimenter dans le contexte d'une communauté réelle (les chercheurs du LIRMM) avec son corpus de connaissances (leurs publications scientifiques).

VIEWPOINTS permet à des *agents*[1] humains ou artificiels d'éliciter leurs points de vue (*viewpoints*) concernant la proximité sémantique entre des *objets* de connaissance du Web, que ce soit des supports (*documents*), des fournisseurs (*agents*), ou des descripteurs (*topics*) de la connaissance. L'approche VIEWPOINTS vise à tirer pleinement parti des points de vue explicites que des pairs, des collègues ou des compagnons ont précédemment exprimés. Un *viewpoint* est émis par un *agent* humain ou artificiel et porte sur la proximité sémantique entre deux objets de connaissance (*agents*, *documents* ou *topics*). Dans cette approche, la dynamique des points de vue individuels à travers les cycles réitérés de recherche d'information et de *feedback* fait émerger la connaissance collective sous forme de chemins renforcés ou atténués reliant les objets de connaissance au sein d'un graphe communautaire, selon la métaphore d'un cerveau dont les circuits neuronaux s'ajustent en continu au monde réel.

Le reste de l'article est organisé de la façon suivante : la section 2 pose le cadre de notre approche et présente notre vision. La section 3 présente le formalisme VIEWPOINTS. La section 4 illustre l'approche et aborde les questions mentionnées ci-dessus en mettant en œuvre un prototype dans un contexte d'utilisation réel : l'indexation des métadonnées d'une base de données bibliographique tirée de HAL-LIRMM. La section 5 discute les avantages et limitations de l'approche. Enfin, la section 6 tire les conclusions de cette première expérimentation pour orienter la poursuite de notre travail.

2 Etat de l'art et vision

L'espace de partage de connaissances où collaborent des agents humains et des agents artificiels peut être considéré comme un espace de consolidation-négociation de « points de vue », voire de « visions du monde » [20], [19], [8]. Dans les domaines où la consolidation est stabilisée apparaissent des ontologies contextuelles, cependant la majeure partie de l'espace reste peuplée de micro-expressions de sémantiques individuelles, avec parfois des éditeurs qui produisent un savoir consolidé en tissant des liens entre les visions du monde de plusieurs experts.

T. Gruber définit l'ontologie comme « la spécification d'une conceptualisation » [6]. Les ontologies sont le plus souvent constituées par un ensemble d'experts à l'issue d'un long processus de convergence de leurs représentations respectives (vision « par le haut »). Il est à noter que les ontologies contextuelles, même mises à jour quotidiennement (e.g., Gene Ontology), peuvent conduire à des situations où le rythme d'entretien du consensus est dépassé par le rythme d'évolution des connaissances de la communauté [15]. Quand le temps de la convergence est trop lent, l'étude des conditions d'émergence prend un intérêt accru ; c'est ainsi que certains travaux [1] sont dédiés à la possibilité de créer des ontologies « par le bas », directement à partir des interactions et de l'évolution d'un système. Toute la question réside alors dans la représentation de ces interactions. Dans la grande diversité des approches

[1] Les termes en italiques dans ce texte sont les termes réservés du formalisme ; ils sont définis dans la section 3.

qui existent aujourd'hui, nous distinguons deux grands courants : celui où l'agent porteur d'une sémantique est représenté explicitement dans le système, et celui où il ne l'est pas.

Pour commencer par le second, l'approche Topics Maps [19] [3] associe des thèmes (topics) à des ressources du Web via des occurrences et des contextes (scope), en mettant en relation des topics et des contextes ; cependant l'agent émetteur est ignoré. L'approche permet d'exploiter des relations topic-topic ou topic-document, ce qui est plus riche que la seule relation topic-document utilisée en recherche d'information. Le réseau de relations ainsi formé est très commode pour la navigation, mais reste difficilement exploitable de façon automatique (en termes d'algorithmes de graphe) de par la variété des types de relations et de leurs arités.

Dans le premier courant citons d'abord l'approche Hypertopic [20] où l'agent est introduit comme contributeur à une vue partielle portant sur une ou plusieurs ressources ; les visions du monde obtenues sont ainsi partagées entre plusieurs acteurs. Citons également Gesche et al. [5] qui propose une approche où les utilisateurs sont émetteurs et contradicteurs de points de vue, mais où aucun formalisme n'est proposé pour faciliter l'exploitation de l'espace de connaissances.

Les réseaux sociaux ou les sites collaboratifs sont bien évidemment les sources d'interactions principales pour ce premier courant. Ce sont des sources où puiser pour obtenir par émergence une sémantique collective à partir de l'expression des sémantiques individuelles. Le tagging social notamment est un mode d'interaction très répandu (e.g., flickr, delicious, last.fm) permettant à partir de micro-contributions d'obtenir et d'entretenir une folksonomie (de la contraction de folk et taxonomie). Les approches [15] [14] [18] considèrent de manière explicite les agents émetteurs de ces tags ; l'association d'un tag à une ressource par un agent étant généralement modélisée par un triplet agent-tag-document. Toutefois même si ces approches considèrent l'agent comme source explicite des associations, aucune ne considère l'agent comme objet potentiel de l'interaction c.-à-d. qu'il n'existe aucun triplet agent-agent-agent ou agent-agent-tag.

Un cas particulier intéressant est celui des systèmes contributifs, avec certaines approches utilisant le jeu comme levier de motivation, que ce soit pour l'enrichissement de folksonomie [9] ou pour la collecte d'informations lexicales [10] ; pour une revue des systèmes de recommandations qui mentionne en particulier les approches collaboratives voir [2]. Parmi les travaux récents qui visent à utiliser les tags d'une communauté pour enrichir un thésaurus, Limpens et al. [13] offrent aux utilisateurs de participer à la structuration d'un réseau de tags en mettant en place un protocole d'interaction impliquant les retours des utilisateurs et en recueillant et confrontant, grâce à des arbitres désignés, leurs points de vue sur les relations entre tags (related, broader, narrower, etc.).

Un élément de formalisation fédérateur du premier courant est la représentation basée sur le triplet *agent-document-topic*, où le terme *topic* englobe à la fois les tags cités ci-dessus et les concepts que l'on trouve dans les ontologies. Un autre élément fédérateur est le calcul de similarités entre ces objets (cf. [14] pour une revue et comparaison des différentes approches). En particulier, Quattrone et al. [17] offre une définition de la similarité basée sur le principe de renforcement mutuel : « deux tags sont considérés similaires s'ils ont été associés à des ressources similaires et vice-et-versa deux ressources sont considérées similaires si elles ont été associées à des tags similaires ». La prise en compte de similarités peut déboucher sur des mesures parfois appelées distances sémantiques [7][11] qui sont utiles pour l'annotation, la désambiguïsation, l'alignement de concepts ou la recherche d'information. Toutefois, si les distances sémantiques sont souvent utilisées pour déterminer la proximité entre *topics*, nous ne connaissons pas d'approche pour déterminer la proximité entre *topic*, *agent* et *document* d'une manière générique et symétrique qui permettrait d'exploiter pleinement le graphe qui sous-tend l'espace de partage de connaissances.

L'approche VIEWPOINTS se situe dans une vision qui cherche à exploiter directement la dynamique des interactions pour produire des connaissances par émergence, avec prise en compte explicite des agents, non seulement en tant que sources de points de vue, mais

également comme objets du graphe de connaissances. Notre but n'est pas de produire des ontologies mais plutôt de faire apparaître dans cet espace des lignes de force qui favorisent la sérendipité et la consolidation. Pour pouvoir à terme puiser dans les immenses volumes de données hétérogènes que sont les réseaux sociaux, nous nous appuyons sur un formalisme minimaliste centré sur une relation unique : le *viewpoint*, qui est défini dans la section 3.

Cette approche VIEWPOINTS s'inscrit dans une tentative de synthèse opérant les choix suivants:

- La représentation des connaissances s'appuie sur un hypergraphe constitué de triplets à la manière décrite dans [18], qui peut aussi être vu comme un graphe biparti avec d'un côté des objets et de l'autre des connecteurs. Dans le cas de VIEWPOINTS, les objets sont les *agents*, les *documents* et les *topics* ; nous traitons ces trois classes d'objets comme sous-classes d'une seule classe regroupant tous les objets de connaissance. Les connecteurs sont les *viewpoints* qui peuvent être des contributions initiales à la connaissance commune ou bien des *feedbacks* suite à une recherche d'information.
- L'accent est mis sur l'émergence au sein du graphe biparti, de même que [1].
- Le moteur évolutif du graphe est basé sur la dynamique « recherche d'information et *feedback* », à la manière de [13], mais sans procédure d'arbitrage.
- Nous définissons une distance métrique sur l'ensemble des objets de connaissance formé par les *agents*, les *documents* et les *topics* (alors que les distances sémantiques que l'on trouve dans la littérature s'appliquent à des sous-classes homogènes) et nous basons le calcul dynamique de cette distance sur l'ensemble des *viewpoints* (contributions initiales ou *feedbacks* des utilisateurs).

3 Le formalisme VIEWPOINTS

Nous présentons ci-dessous le formalisme VIEWPOINTS qui a déjà fait l'objet d'une description dans [12]. Considérons une collection O d'objets, comportant trois sous-classes A, D et T :

- les objets de la sous-classe A sont interprétés comme des *agents*. Les *agents* fournissent les point de vues ; ils sont soit humains (émetteurs de points de vue) soit artificiels (par exemple, les extracteurs de *topics*).
- les objets de la sous-classe D sont interprétés comme des *documents*. La notion de *document* unifie tous les supports de connaissance (textes, cartes, vidéos, etc.).
- les objets de la sous-classe T sont interprétés comme des *topics*. Le concept de *topic* unifie tous les moyens de description des *documents* ou des *agents* (mots clés, taxon, etc.), ou les thèmes de réflexion (sujets de discussion dans les fora).

Soit W l'ensemble de tous les couples constitués d'un *agent* de A et d'une paire d'objets quelconques de O ; les éléments de W sont de la forme $(a_i, \{o_j, o_k\})$ avec $a_i \in A$ et $o_j, o_k \in O$. En notant $a_i \rightarrow \{o_j, o_k\}$ l'élément $w = (a_i, \{o_j, o_k\})$, nous obtenons six formes de base : $a_1 \rightarrow \{d_1, t_1\}$, $a_1 \rightarrow \{t_1, t_2\}$, $a_1 \rightarrow \{d_1, d_2\}$, $a_1 \rightarrow \{a_2, t_1\}$, $a_1 \rightarrow \{a_2, d_1\}$ et $a_1 \rightarrow \{a_2, a_3\}$.

Un *viewpoint* est alors défini comme un triplet $v = (w, \alpha, \tau)$ où $w \in W$ et $\alpha, \tau \in \mathbb{R}$:

1. $w = a_1 \rightarrow \{o_2, o_3\}$ s'interprète de la façon suivante : « l'*agent* a_1 déclare une proximité entre les deux objets o_2 et o_3 ».
2. α est l'évaluation de cette proximité entre o_2 et o_3 telle qu'elle est donnée par l'*agent* a_1 à l'instant τ ; $\alpha \geq 0$.

Un exemple de *viewpoint* est l'association : « selon l'*agent* a_1, le *document* d_1 est pertinent relativement au *topic* t_1 à l'instant τ » ; ce *viewpoint* est formalisé en gardant des rôles symétriques pour d_1 et t_1 : $v = (a_1 \rightarrow \{d_1, t_1\}, +1, \tau)$.

Le *Knowledge Graph* (KG) est le graphe biparti suivant :

- les sommets de KG sont les éléments de $O \cup W$.
- les arêtes de KG sont obtenues à partir des éléments de W : chaque $w = a_1 \rightarrow \{o_2, o_3\}$ produit 3 arcs orientés : $a_1 \rightarrow w$, $w \rightarrow o_2$ and $w \rightarrow o_3$.
- les sommets pris dans W sont étiquetés par (α, τ) ; ce sont les *viewpoints.*

Les *viewpoints* de KG vont servir de base à la définition d'une distance sur O×O :

- soit $v = (w, \alpha, \tau) = (a_1 \rightarrow \{o_2, o_3\}, \alpha, \tau)$ un *viewpoint* connectant $\{o_2, o_3\}$, on écrit : $jump_v(\{o_2, o_3\}) = \alpha$ pour exprimer cette connexion élémentaire.[2]
- soit $W\{o_2, o_3\} = \{w \in W \mid \exists a_i \in A, (a_i, \{o_2, o_3\})\}$, l'ensemble des *viewpoints* connectant o_2 et o_3.
- l'ensemble des connexions entre deux objets $\{o_2, o_3\}$ dues aux différents *agents* constitue un « lien de proximité » nommé *synapse* dont on peut calculer la force en faisant la somme algébrique de tous les *jumps* reliant ces deux objets. On obtient une valeur positive:

$$synapse(\{o_2, o_3\}) = \sum_{W\{o_2, o_3\}} jump_v(\{o_2, o_3\})$$

- il est alors possible de considérer le graphe non orienté construit sur les sommets O reliés par les *synapses*, et de définir une distance métrique à partir des plus courts chemins dans ce graphe. Nous appelons *ψ-distance* cette distance métrique construite sur O×O.
- le *m-neighborhood* d'un objet 'o_q' de O est alors l'ensemble des objets 'o' ∈ O vérifiant : *ψ-distance* $(o_q, o) \leq m$. Le calcul du *m-neighborhood* s'inspire de l'algorithme du plus court chemin de Dijkstra en utilisant le paramètre m pour limiter la propagation dans le graphe. Ainsi, pour un objet 'o_q', *m-neighborhood*(o_q) renvoie les objets appartenant aux chemins partant de o_q et de longueur inférieure ou égale à m.

Dans KG, la dynamique des *viewpoints* est le reflet direct de la consolidation et de la confrontation des opinions individuelles au sein de la communauté. Toute recherche d'information s'appuie sur le calcul du *m-neighborhood*, lui-même basé sur l'ensemble des *viewpoints* présents au moment de la recherche. En retour, l'*agent* qui avait émis la requête est invité à évaluer la proximité entre l'objet initialement recherché et les objets du *m-voisinage* qui lui sont proposés. Ces « feedbacks » produisent des *viewpoints* nouveaux qui influenceront les prochains calculs de *ψ-distance*. Il y a donc coévolution des *synapses* au rythme des interactions communautaires.

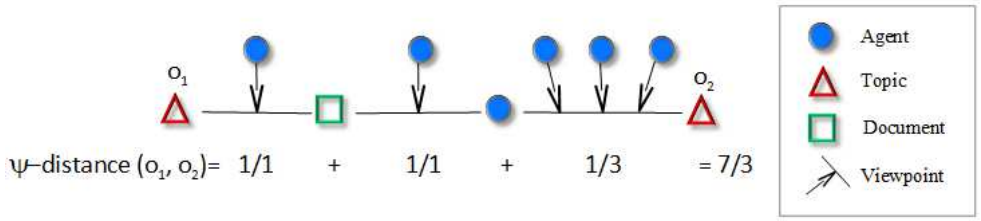

FIGURE 1 – Illustration du calcul de la distance entre deux objets de KG.

En outre, grâce au nouveau graphe construit à partir de O en évaluant toutes les *synapses* de KG, il devient possible de suivre l'évolution de trois structures émergentes : les réseaux de *documents* (bibliographies), les réseaux d'*agents* (sociogrammes), les réseaux de *topics* (onto-terminologies) comme illustré sur la figure 2. Cette analyse sera approfondie dans une prochaine publication.

[2] Dans la pratique, pour un instant τ donné, le graphe KG contiendra un seul *viewpoint* $(a_1 \rightarrow \{o_2, o_3\}, \alpha, \tau)$; la notation $jump_v$ est préférée à la notation $jump_a$ car l'agent a_1 peut émettre plusieurs *viewpoints* sur $\{o_2, o_3\}$ à des instants différents.

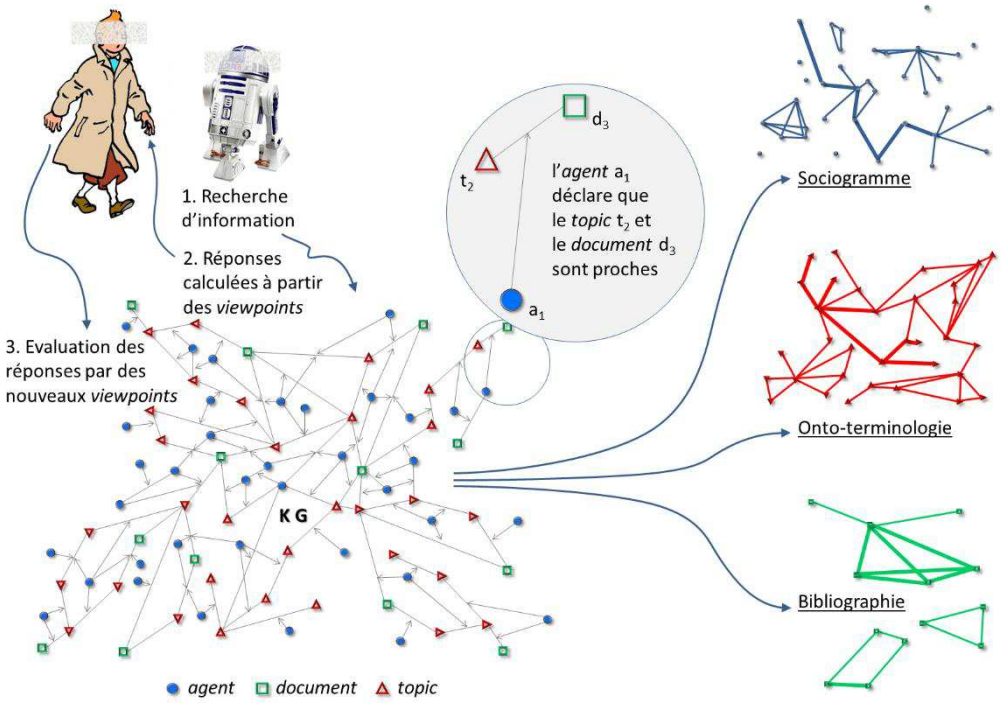

FIGURE 2 – Vue d'ensemble de l'approche VIEWPOINTS (toute ressemblance avec des *agents* existant ou ayant existé est purement fortuite). Par souci de lisibilité, α et τ ne sont pas représentés dans les *viewpoints*. Dans la partie droite, les traits représentent des synapses et leur épaisseur en représente la force.

4 Illustration de l'approche dans un contexte réel de recherche d'information

4.1 Ressource utilisée

Pour construire une application et tester cette approche sur des données réelles, nous avons choisi les ressources bibliographiques du HAL-LIRMM[3] comme corpus de données à indexer avec VIEWPOINTS. C'est une base de données de toutes les publications du LIRMM. Notre choix s'est porté vers cette ressource car :

- Chaque document est accompagné de métadonnées telles que les auteurs et les mots-clés choisis par ces auteurs pour décrire leurs publications. Cela en fait un jeu de données approprié pour illustrer le potentiel du formalisme.
- Nous avons accès à ce jeu de données de taille raisonnable.
- Ces données concernent nos collègues et notre laboratoire ce qui est donc pertinent pour évaluer/calibrer l'approche VIEWPOINTS et motiver nos collègues à fournir leur évaluation et leurs *viewpoints* lors du *feedback*.

4.2 Modèle adopté pour le graphe de connaissances

Dans un souci de comparaison et d'alignement avec le moteur de recherche fourni par HAL-LIRMM, nous avons sélectionné les métadonnées les plus simples pour l'initialisation du graphe de connaissance : pour un *document* d, pour chaque auteur a et pour chaque mot-clé t nous créons un *viewpoint* (a→{d, t}, +1, 0). Cette procédure est répétée sur chaque *document*. Dans notre application, basée sur les données de septembre 2013, 1663 *agents*, 5219 *documents* et 5846 *topics* nous donnent 42860 *viewpoints*.

[3] http://hal-lirmm.ccsd.cnrs.fr

Dans ce modèle, pour mettre en évidence l'impact des contributions, nous affectons un poids (α=3) aux viewpoints générés par les utilisateurs lors du *feedback* supérieur au poids des viewpoints créés lors de l'initialisation du graphe de connaissance avec les métadonnées.

Le prototype est implémenté en Java. Nous avons utilisé l'API d'affichage de graphe JUNG[4] afin d'avoir une visualisation du graphe de connaissance et des plus courts chemins.

4.3 Illustration de l'utilisation du prototype et du graphe de connaissance

La figure 3 illustre un sous-ensemble des objets au voisinage du *topic* 'Semantic Web'.[5] Une recherche de 'Semantic Web' sur HAL-LIRMM ne retourne que les objets qui sont directement liés à cette requête, c.-à-d. les articles ayant 'Semantic Web' dans leurs mots-clés ainsi que les auteurs de ces articles. Cependant, dans notre prototype, grâce à l'utilisation de la *ψ-distance*, la requête 'Semantic Web' renvoie en une seule recherche le *m-neighborhood* de ce *topic*, c.-à-d. tous les objets pour lesquels il existe un chemin de longueur inférieure ou égale à 'm' vers ce *topic*; par exemple 'Knowledge Management' ou 'Data Linking'.

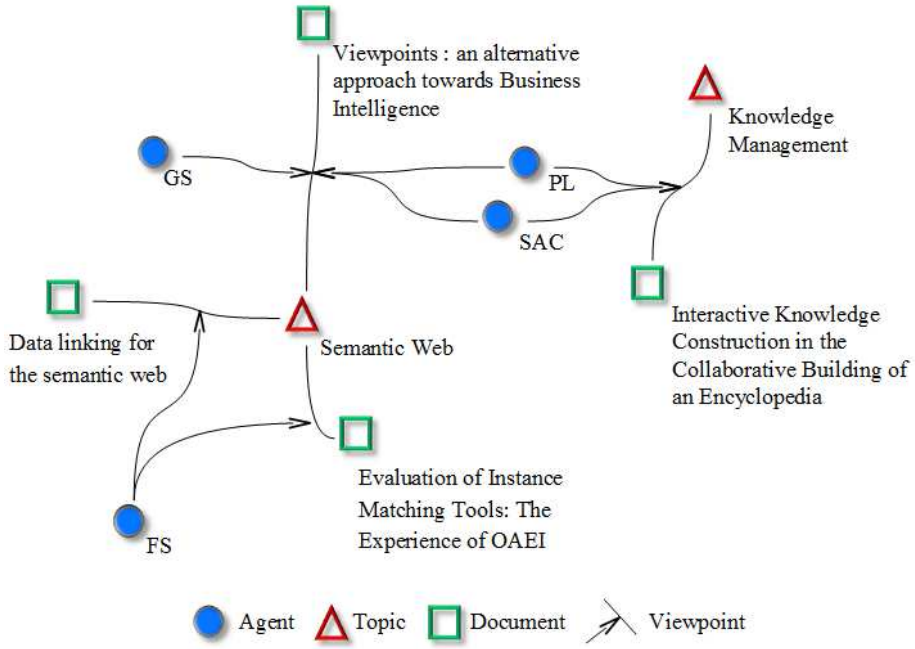

FIGURE 3 – Sous graphe d'objets au voisinage de 'Semantic Web' extraits à partir de KG.

Le cas traité pour illustrer l'approche se déroule en trois étapes :

Étape 1 : Guillaume Surroca (GS) exécute une recherche sur le *topic* 'Knowledge Management'.[6] L'interface présente alors les résultats dans trois onglets ('Documents', 'Agents', 'Topics') comme illustré dans la figure 4. L'*agent* Francois Scharffe (FS) y figure à une distance de 0,58 de l'objet recherché. Le prototype donne à l'utilisateur l'explication des résultats qu'il propose : en effet, l'utilisateur peut visualiser pour chaque résultat un des plus courts chemins reliant l'objet de la requête et le résultat dans le graphe de connaissances (bouton 'Path'). De plus, l'utilisateur peut valider chaque résultat (boutons 'Right') et émettre ainsi en guise de *feedback* de nouveaux *viewpoints* qui viendront nourrir le graphe pour les prochaines requêtes comme illustré dans la suite du scénario.

[4] Java Universal Network/Graph Framework, projet financé par la NSF américaine : http://jung.sourceforge.net

[5] Pour garder les schémas lisibles nous affichons dans le graphe de connaissances seulement les nœuds servant à l'illustration.

[6] Dans le prototype, un utilisateur saisit une chaine de caractères qui permet d'identifier l'objet de la requête (document, agent ou topic) par auto-complétion c'est-à-dire en se limitant explicitement aux objets de connaissance déjà présents dans le graphe.

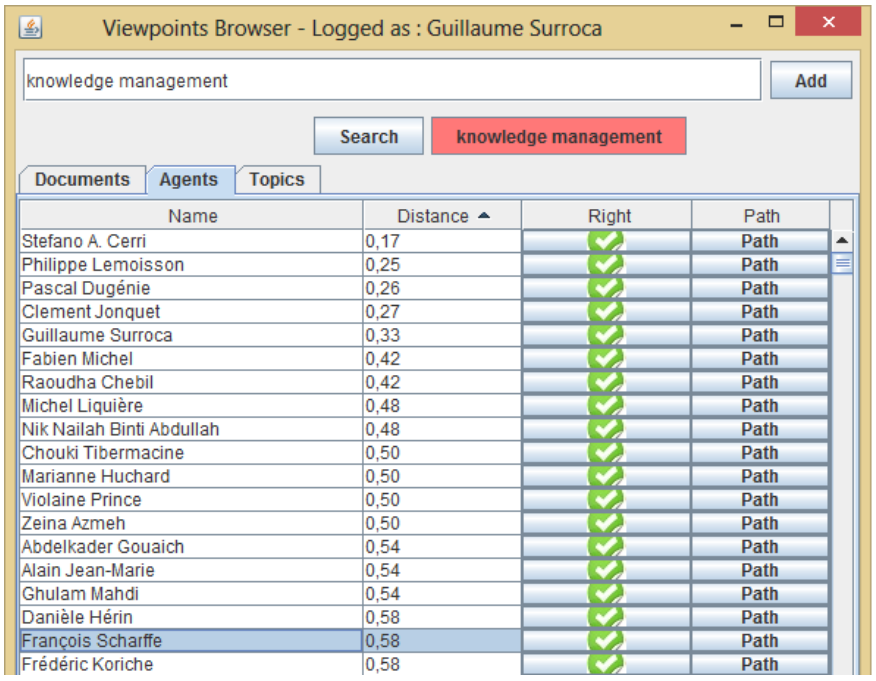

FIGURE 4 – Illustration d'une recherche sur 'Knowledge Management' dans l'interface. Le nom de l'utilisateur connecté apparait et permettra d'identifier l'émetteur des *viewpoints* lors du *feedback*.

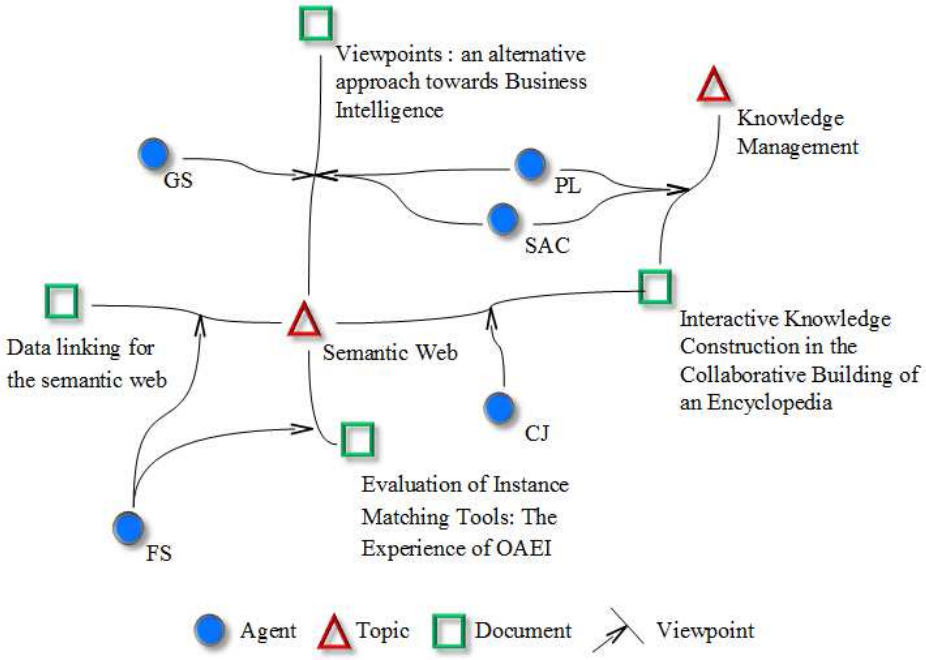

FIGURE 5 – Impact du *feedback* de CJ sur le graphe de connaissances.

Étape 2 : Ensuite, Clément Jonquet (CJ) fait une recherche sur le *topic* 'Semantic Web', et obtient comme résultat l'article 'Interactive Knowledge Construction in the Collaborative Building of an Encyclopedia' (IKC) ; son *feedback* consiste à approuver le résultat en émettant un nouveau *viewpoint* reliant ce *document* au *topic* 'Semantic Web'. La figure 5 illustre le graphe de connaissances après la contribution de CJ. Ce nouveau *viewpoint*, de poids $\alpha=3$ contribue à une *synapse* (IKC, Semantic Web) plus forte que les *synapses*

précédentes (SAC, Semantic Web) ou (PL, Semantic Web) ; ainsi il existe un nouveau plus court chemin et la distance diminue.

Étape 3 : Finalement, GS refait une recherche (figure 6) sur 'Knowledge Management', et cette fois-ci l'*agent* FS apparaîtra plus haut dans la liste des résultats (ordonnés par distances) étant donné que ces deux objets se sont rapprochés (0.50 < 0.58).

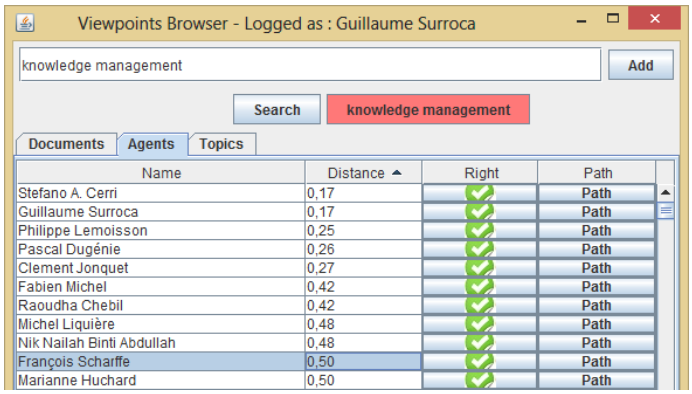

FIGURE 6 – Impact du *feedback* de CJ sur la recherche.

VIEWPOINTS offre un potentiel accru pour la recherche d'information :

- Pour un *agent* donné, le prototype retourne les *agents* au voisinage permettant d'identifier d'autres contributeurs ou collaborateurs potentiels. Il permet en outre d'identifier les *documents* proches de cet *agent* (sans se limiter aux publications dont il est auteur) et ses *topics* d'intérêts explicites (mot clés de ses publications) ou implicites (mots clés d'autres publications dont il est proche).
- Pour un *topic* donné, le prototype permet non seulement d'identifier les *documents* pertinents (comme n'importe quel moteur de recherche par mot clé) mais permet également d'identifier les experts pour ce *topic* et les *topics* proches dans le graphe de connaissance (illustration de la terminologie spécifique à une base de connaissances).
- Pour un *document* donné, un utilisateur peut trouver d'autres *documents* similaires (en plus des *topics* et *agents* proches).

5 Discussions

5.1 Aspects liés au corpus de connaissance et à la dynamique des contributions

Dans cette illustration de notre approche, le graphe de connaissances est obtenu par extraction d'une unique source de données ; il sera intéressant de l'enrichir par d'autres informations telles que les projets de recherche du LIRMM, l'organisation structurelle du laboratoire (équipes, membres et thématiques), etc. Ces autres jeux de données, une fois traduits en *viewpoints*, contribueront aux distances entre objets et participeront à la structuration du graphe de connaissances.

Par ailleurs, il faut noter que le graphe présenté est le graphe initial obtenu par extraction des métadonnées (à l'exception du *viewpoint* rajouté par CJ). C'est seulement après un nombre de requêtes et *feedbacks* suffisant que des lignes de force matérialisées par les synapses entre *viewpoints* émergeront : les *feedbacks* créent ou renforcent les synapses dans le graphe de connaissance. Par l'intermédiaire des calculs de voisinages, les *feedbacks* impactent donc directement les futures recherches d'autres utilisateurs, ce qui permet de parler d'intelligence collective.

Cette discussion amène la prise de conscience d'un challenge majeur dans notre approche : en utilisant les point de vues des utilisateurs comme contributions et sources de l'indexation

des objets de connaissance le problème de l'indexation est transféré vers d'autres problème : (i) comment motiver les contributions et (ii) comment tirer le maximum de ces contributions ?

5.2 Aspects liés aux choix de modélisation (façon dont nous avons transcrit le corpus)

Transformer des métadonnées en *viewpoints* suppose un ensemble de choix de modélisation. Ainsi, le choix exposé section 4.2 : « dans un *document* d, pour chaque auteur a et pour chaque mot-clé t nous créons un *viewpoint* (a→{d, t}, +1, 0) » est celui d'un modèle simple et immédiat. Il aurait été possible de rajouter par exemple des *viewpoints* exprimant de manière accentuée la paternité des *documents* (a→{a, d}, +10, 0), de façon à mettre en œuvre un modèle plus expressif. La détermination d'un modèle ayant un bon rapport expressivité/simplicité et son calibrage sont les prochaines étapes de notre travail sur ce prototype au moyen de poids différents et de typage des *viewpoints*.

En outre, étant donné que les *topics* sont pour le moment des mots-clés librement choisis par les auteurs lorsqu'ils ont enregistré leurs publications sur HAL-LIRMM le graphe de connaissance connaît les problèmes classiques des folksonomies (ambiguïté, polysémie, multilinguisme, etc.). Par exemple, certains *topics* peuvent être ambigus et créer de faux chemins ; ils peuvent également représenter le même concept. Une stratégie de remédiation serait l'intégration d'un *agent* artificiel exploitant une ontologie suffisamment riche pour éliminer les ambiguïtés et s'appuyant sur celle-ci pour exprimer sous forme de *viewpoints* des proximités sémantiques précises. Il est à noter que ce problème lié aux folksonomies n'existe pas avec des sources telles que PubMed (publications biomédicales) respectant une terminologie standardisée (MeSH).

Cette discussion amène des questions plus générales : comment extraire sous forme de *viewpoints* des données ou métadonnées explicites formalisées dans des bases de connaissances, des jeux de données, ou des ontologies ? Comment extraire des *viewpoints* quand ils sont implicites (e.g., exprimés sous forme textuelle) ? Comment lier nos *topics* aux schémas du Web sémantique (vocabulaires et ontologies existantes) ?

5.3 Aspects liés au formalisme lui même

Le fait d'obtenir dans les résultats d'une recherche des objets qui ne sont pas directement liés (par les métadonnées) à l'objet de la requête est la principale source de sérendipité. Il est dans la nature de la *ψ-distance* d'ouvrir des chemins par transitivité : ainsi, dans le prototype, deux *topics* attachés au même *document* deviennent automatiquement « un peu proches », ce qui peut ne pas toujours être approprié. Il ne faut pas oublier cependant que c'est la dynamique de l'interaction qui produit les liens sémantiques les plus forts : une proximité discutable doit pouvoir être remise en question par des *viewpoints*. La réflexion concernant une gestion discriminante des chemins hétérogènes (*topic-document-topic* par exemple) par l'algorithme de plus court chemin est en cours.

5.4 La question de l'évaluation

Que ce soit pour perfectionner le formalisme, pour choisir et calibrer un modèle, ou pour établir la preuve de concept dans un scénario réel, la question de l'évaluation du gain en intelligence collective est importante et difficile. Il s'agit en effet d'évaluer un apprentissage collectif, sans objectifs initiaux. Nous avons des pistes pour cet aspect : (i) l'observation de l'évolution de la structure du graphe de connaissances au fil des interactions ; (ii) l'observation qualitative du flux de *viewpoints*. La réflexion est en cours pour trouver des benchmarks et un protocole de simulation satisfaisants. Des benchmarks de recherche d'information nous permettront d'évaluer (en termes de précision/rappel) le scénario de recherche *topic-document*. Nous prévoyons également une évaluation en situation réelle par les utilisateurs.

6 Conclusions et perspectives

En traitant le corpus des publications des chercheurs du LIRMM, nous avons montré l'opérationnalité de l'approche VIEWPOINTS dans un contexte de recherche d'information, et nous avons apporté des éléments de réponse aux questions énoncées dans l'introduction :

- Le graphe de connaissances réifie en toute transparence la sémantique collective de la communauté. Il contient toutes les explications concernant l'émergence de cette sémantique à partir des contributions. Le processus de réponse aux requêtes est donc lui aussi transparent, permettant à l'utilisateur de trouver des documents ou des données dignes de confiance sur ses sujets d'intérêt.
- La géographie de la connaissance ainsi produite permet aisément de trouver les bonnes personnes pour échanger, argumenter et capitaliser sur un sujet particulier.
- Au fil des interactions, la structure du graphe élicite une proximité sémantique entre certains *topics*, en reflétant les points de vue des membres de la communauté. Ceci est un premier pas vers des processus d'agrégation susceptibles de faire émerger de nouvelles connaissances.

Ce premier prototype nous a permis d'expérimenter le calcul de la *ψ-distance* sur des données réelles, mais surtout de valider l'aptitude du formalisme à supporter un modèle traitant la recherche d'information scientifique. Pour achever la preuve de concept, nous envisageons deux scénarios d'expérimentation à plus grande échelle afin de tenter d'appréhender le gain en intelligence collective :

- Un scénario centré sur les ontologies et les ressources de données biomédicales où l'objectif sera d'intégrer sous forme de *viewpoints* : (i) des annotations (manuelles ou automatiques) basées sur les ontologies ou (ii) des données structurées comme des données liées sur le Web. L'objectif sera de montrer les apports de l'approche pour l'indexation sémantique.
- Un scénario centré sur une thématique environnementale, où l'objectif sera d'exploiter les connaissances explicites et implicites des chercheurs du Cirad pour faire émerger une connaissance communautaire dans une dynamique favorisant la controverse.

La question de l'évaluation évoquée section 5 sera abordée simultanément à la définition de ces scénarios. Ces scénarios nous permettront de poursuivre les travaux sur le formalisme et sur l'exploitation des graphes de connaissances en bénéficiant des retours des utilisateurs. Nous envisageons l'exploitation de la dimension temporelle, simultanément avec des études de clusterisation, pour observer l'évolution des graphes et étudier les dynamiques d'évolution des connaissances au sein des communautés.

Remerciements

Ce travail a bénéficié des soutiens du Cirad et du projet SIFR (Semantic Indexing of French Biomedical Resources) financé en partie par le programme JCJC de l'Agence nationale de la Recherche (ANR-12-JS02-01001), l'Université Montpellier 2, le CNRS et l'Institut de Biologie Computationnelle de Montpellier.

Références

[1] ABERER, K., CUDRE-MAUROUX, P., OUKSEL, A., CATARCI, T., HACID, M.-S., ILLARRAMENDI, A., KASHYAP, V., MECELLA, M., MENA, E., NEUHOLD, E., TROYER, O., RISSE, T., SCANNAPIECO, M., SALTOR, F., SANTIS, L., SPACCAPIETRA, S., STAAB, S., AND STUDER, R. Emergent Semantics Principles and Issues. In *Database Systems for Advanced Applications*, Y. Lee, J. Li, K.-Y. Whang, and D. Lee, Eds., vol. 2973 of *Lecture Notes in Computer Science*. Springer, 2004, pp. 25–38.

[2] ADOMAVICIUS, G., AND TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering 17*, 6 (June 2005), 734–749.

[3] CAUSSANEL, J., CAHIER, J.-P., ZACKLAD, M., AND CHARLET, J. Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ? In *13èmes journées francophones d'Ingénierie des Connaissances, IC'02* (Rouen, France, May 2002), p. 12.

[4] GANDON, F., FARON-ZUCKER, C., AND CORBY, O. *Le web sémantique - Comment lier les données et les schémas sur le web ?* Dunod, 2012.

[5] GESCHE, S., CAPLAT, G., AND CALABRETTO, S. Managing Difference of Opinion in Semantic Structures. In *International Workshop on Semantically Aware Document Processing and Indexing, SADPI'07* (Montpellier, France, May 2007), H. Betaille, J.-Y. Delort, M.-L. Mugnier, J. Nanard, and M. Nanard, Eds., ACM, pp. 79–86.

[6] GRUBER, T. R. A translation approach to portable ontologies. *Knowledge Acquisition 5*, 2 (June 1993), 199–220.

[7] HARISPE, S., RANWEZ, S., JANAQI, S., AND MONTMAIN, J. The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics* (October 2013).

[8] IACOVELLA, A., BENEL, A., PETARD, X., AND HELLY, B. *La redocumentarisation du monde*. Cépaduès, 2006, ch. Corpus scientifiques numérisés : Savoirs de référence et points de vue des experts, pp. 117–130.

[9] KRAUSE, M., AND ARAS, H. Playful tagging: folksonomy generation using online games. In *18th International Conference on World Wide Web, WWW'09* (Madrid, Spain, 2009), pp. 1207–1208.

[10] LAFOURCADE, M. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *7th International Symposium on Natural Language Processing, SNLP'07* (Pattaya, Chonburi, Thailand, December 2007), p. 7.

[11] LEE, W.-N., SHAH, N. H., SUNDLASS, K., AND MUSEN, M. A. Comparison of Ontology-based Semantic-Similarity Measures. In *American Medical Informatics Association Annual Symposium, AMIA'08* (Washington DC, USA, November 2008), pp. 384–388.

[12] LEMOISSON, P., SURROCA, G., AND CERRI, S. A. Viewpoints: An Alternative Approach toward Business Intelligence. In *eChallenges e-2013 Conference* (Dublin, Irland, October 2013), p. 8.

[13] LIMPENS, F., GANDON, F., AND BUFFA, M. Un cycle de vie complet pour l'enrichissement sémantique des folksonomies. In *11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, EGC'11* (Brest, France, Janvier 2011), A. Khenchaf and P. Poncelet, Eds., vol. E-20 of *Revue des Nouvelles Technologies de l'Information*, Hermann, pp. 389–400.

[14] MARKINES, B., CATTUTO, C., MENCZER, F., BENZ, D., HOTHO, A., AND STUMME, G. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *18th International Conference on World Wide Web, WWW'09* (Madrid, Spain, 2009), pp. 641–650.

[15] MIKA, P. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In *4th International Semantic Web Conference, ISWC'05* (Galway, Ireland, November 2005), Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, Eds., vol. 3729 of *Lecture Notes in Computer Science*, Springer, pp. 522–536.

[16] O'REILLY, T. What Is Web 2.0. Oreilly's blog post, September 2005.

[17] QUATTRONE, G., FERRARA, E., MEO, P. D., AND CAPRA, L. Measuring Similarity in Large-scale Folksonomies. In *23rd International Conference on Software Engineering and Knowledge Engineering, SEKE'11* (Miami Beach, FL, USA, July 2011), pp. 385–391.

[18] SPECIA, L., AND MOTTA, E. Integrating Folksonomies with the Semantic Web. In *4th European Semantic Web Conference, ESWC'07* (Innsbruck, Austria, June 2007), E. Franconi, M. Kifer, and W. May, Eds., vol. 4519 of *Lecture Notes in Computer Science*, Springer, pp. 624–639.

[19] STEVE PEPPER, G. O. G. Towards a General Theory of Scope. In *Extreme Markup Languages Conference* (Montreal, Canada, August 2001), p. 4.

[20] ZACKLAD, M., BENEL, A., ZAHER, L., LEJEUNE, C., CAHIER, J.-P., AND ZHOU, C. Hypertopic: une métasémiotique et un protocole pour le Web socio-sémantique. In *18ème journées francophones d'Ingénierie des Connaissances, IC'07* (Grenoble, France, July 2007), F. Trichet, Ed., Cépaduès, pp. 217–228.